

Major changes in the core developmental pathways of nematodes: *Romanomermis culicivorax* reveals the derived status of the *Caenorhabditis elegans* model

Philipp H. Schiffer^{1†}, Michael Kroiher¹, Christopher Kraus¹, Georgios D. Koutsovoulos², Sujai Kumar², Julia I. R. Camps¹, Ndifon A. Nsah¹, Dominik Stappert³, Krystalynne Morris⁴, Peter Heger¹, Janine Altmüller⁵, Peter Frommolt⁵, Peter Nürnberg⁵, W. Kelley Thomas⁴, Mark L. Blaxter² and Einhard Schierenberg¹

[†] corresponding author: PHS - ORCID:0000-0001-6776-0934 - p.schiffer@uni-koeln.de

1. Zoologisches Institut, Universität zu Köln, Cologne, NRW, Germany
2. Institute of Evolutionary Biology, School of Biological Sciences, The University of Edinburgh, Edinburgh, Scotland, UK
3. Institute für Entwicklungsbiologie, Universität zu Köln, Cologne, NRW, Germany
4. Hubbard Center for Genome Studies, University of New Hampshire, Durham, NH, USA
5. Cologne Center for Genomics, Universität zu Köln, Cologne, NRW, Germany

Keywords: nematode, genome, Mermithida, development, *Caenorhabditis*

Abstract

Background

Despite its status as a model organism, the development of *Caenorhabditis elegans* is not necessarily archetypical for nematodes. The phylum Nematoda is divided into the Chromadorea (includes *C. elegans*) and the Enoplea. Compared to *C. elegans*, enoplean nematodes have very different patterns of cell division and determination. Embryogenesis of the enoplean *Romanomermis culicivorax* has been studied in great detail, but the genetic circuitry underpinning development in this species is unknown.

Results

We created a draft genome of *R. culicivorax* and compared its developmental gene content with that of two nematodes, *C. elegans* and *Trichinella spiralis* (another enoplean), and a representative arthropod *Tribolium castaneum*. This genome evidence shows that *R. culicivorax* retains components of the conserved metazoan developmental toolkit lost in *C. elegans*. *T. spiralis* has independently lost even more of the toolkit than has *C. elegans*. However, the *C. elegans* toolkit is not simply depauperate, as many genes essential for embryogenesis in *C. elegans* are unique to this lineage, or have only extremely divergent homologues in *R. culicivorax* and *T. spiralis*. These data imply fundamental differences in the genetic programmes for early cell specification, inductive interactions, vulva formation and sex determination.

Conclusions

Thus nematodes, despite their apparent phylum-wide morphological conservatism, have evolved major differences in the molecular logic of their development. *R. culicivorax* serves as a tractable, contrasting model to *C. elegans* for understanding how divergent genomic and thus regulatory backgrounds can generate a conserved phenotype. The availability of the draft genome will promote use of *R. culicivorax* as a research model.

Background

Species in the phylum Nematoda have a generally conserved body plan. The classic nematode form is dictated by the presence of a hydroskeleton, where longitudinal muscles act against an inextensible extracellular cuticle. What is more surprising is the conservation of organ systems between nematode species, with, for example, the nervous system and the somatic gonad and vulva having very similar external and cellular morphologies. It might be thought that these similar morphologies and cellular structures arise from highly stereotypical developmental programmes, but observational data are emerging that challenge this "all nematodes are equal" view. The embryonic development of the nematode *Caenorhabditis elegans* (Rhabditina, Rhabditida, Chromadorea; see De Ley and Blaxter [1]) has become a paradigmatic model for studying developmental processes in animals, including earliest soma-germline separation, fate specification through cell-cell interactions, and differentiation. The particular mode of development of *C. elegans* is distinct within the major metazoan model organisms, but much of the regulatory logic of its development is comparable to that in other phyla. One key aspect in which *C. elegans* differs from vertebrate and arthropod models is that *C. elegans* has a strictly determined developmental programme [2], with a largely invariant cell lineage giving rise to predictable sets of differentiated cells [3]. Inductive cell-cell interactions are, nevertheless, essential for its correct development [2]. The first description of the early embryonic cell lineage of a nematode, that of *Ascaris* (Spirurina) in the 1880's [4, 5], conforms to the *C. elegans* model.

Early development across all three suborders of the Rhabditida (i.e. Rhabditina, Tylenchina and Spirurina sensu De Ley and Blaxter [1]) is very similar [6, 7]. In general only relatively minor variations on the division pattern observed in *C. elegans*, including heterochrony in the timing of particular cell divisions, and restrictions in cell-cell interaction due to different placement of embryonic blastomeres within the eggshell following altered orientations of cell division spindles have been described in these nematodes [8, 9]. From this large body of work it might be assumed that all nematodes follow a *C. elegans*-like pattern of development. Deviations from the *C. elegans* pattern observed in these rhabditid nematodes indicate that the determined mode of development is subject to evolutionary change, and have assisted in revealing the underpinning regulatory logic of the system. Indeed, a greater role for regulative interactions in early development has been characterised in some rhabditids, such as *Acrobeloides nanus* (Tylenchina) [10, 11]. Regulative development is common in Metazoa, and is also observed in other ecdysozoan taxa (e.g. within the Arthropoda). The determined mode found in *C. elegans* is thus likely to be derived. Molecular and morphological systematics of the phylum Nematoda identify two classes: Chromadorea (including Rhabditida), and Enoplea (subdivided into Dorylaimia and Enoplia) [1, 12] (Figure 1). In several Enoplea,

early embryos do not display polarised early divisions, and observational and experimental evidence argues against a strongly determined mode of development [13, 14]. Strongly determinative development may thus be derived even within Nematoda [15]. This implies that the underpinning developmental system in Nematoda has changed, while maintaining a very similar organismal output. This phenomenon, termed 'developmental system drift' [16], allows independent selection on the mechanism and the final form produced by it. To explore mechanistic aspects of development of enoplean and other non-rhabditid nematodes requires tractable experimental systems with a wealth of underpinning methodological tools and extensive genetic data. While *C. elegans* and its embryos are relatively easily manipulated and observed, and the *C. elegans* genome has been fully sequenced [17], embryos from taxa in Enoplia and Dorylaimia are much harder to culture and manipulate. Few viable laboratory cultures exist and obtaining large numbers of embryos from wild material is difficult. Functional molecular analyses of species in most nematodes, and Enoplia and Dorylaimia in particular, is further hindered by the lack of genetic tools such as mutant analysis or gene-knockdown via RNAi.

While realisation of extensive programmes of comparative experimental embryology across the phylum Nematoda remains a distant research goal, we have taken a parallel genome-based approach. Using the background knowledge of pathways and modules used in other taxa, the underpinning logic of a species' developmental system can be inferred from its genome, and the developmental toolkits of different species can be compared. These comparisons can pinpoint changes in developmental logic between taxa by identifying genes unique to one species or group, and gene losses during evolution, that must result in changed pathway functioning. Efficient generation of genomic resources for non-model species, and the inference of developmental regulatory pathways from the encoded gene sets, is now possible. The majority of the 11 genome sequences determined to date for Nematoda has been from Rhabditida (e.g. *C. elegans* and congeners) [18–25]. A single member of Enoplea, the mammalian parasite *Trichinella spiralis* (Dorylaimia; order Trichocephalida) has been sequenced [26]. *T. spiralis* is ovoviviparous, and proper development requires the intrauterine environment. *T. spiralis* blastomeres are extremely transparent [27] such that individual nuclei are hard to identify (E.S., unpublished observations). Hence this species is of very limited value for image analysis and experimental investigations correlating cellular aspects and the underpinning molecular logic of early development. The genomes of many additional nematode species are being sequenced and annotated [28, 29], but even in this wider sampling of the phylum, Enoplia and Dorylaimia are neglected.

Romanomermis culicivorax (order Mermithida within Dorylaimia), has been established in culture for

decades. *R. culicivora*x infects and kills the larvae of many different mosquito species [30], and is the subject of research programmes investigating its potential as a biocontrol agent of malaria and other disease vectors [30, 31]. *R. culicivora*x and *T. spiralis* differ fundamentally in many life-cycle and phenotypic characters. Free living *R. culicivora*x juveniles actively seek and invade mosquito larvae in the water [32], while *T. spiralis* is transmitted as an arrested, first stage larva encysted in muscle tissue [33]. *R. culicivora*x embryos are easily studied under laboratory conditions, and a single female can produce more than a thousand eggs in culture. They display a developmental pattern that differs markedly from the *C. elegans* model. As in other Dorylaimia and Enoplea [14, 34] the first division is equal, and not asymmetric as in *C. elegans*. *R. culicivora*x also shows an inversion of dorso-ventral axis polarity compared to *C. elegans*. A predominantly monoclonal fate distribution in *R. culicivora*x somatic founder cells indicates fewer modifying inductions between blastomeres [34, 35]. Generation of the hypodermis involves repetitive cell elements extending from posterior to anterior over the remainder of the embryo, a system very different from that of *C. elegans* [35]. In the context of this distinct developmental mode in *R. culicivora*x, we decided to catalogue its developmental toolkit by sequencing the genome, and here present a draft assembly and annotation. We contrast the toolkits identified in *R. culicivora*x and *T. spiralis* with that of *C. elegans*, and of other metazoa, notably the arthropod *Tribolium castaneum*. We conclude that major changes in the regulatory logic of development have occurred during the evolution of nematodes, possibly as a consequence of developmental system drift, and that the model species *C. elegans* represents an extreme derivation from a shared metazoan ground system.

Results and Discussion

Romanomermis culicivorax has a large and repetitive genome

A draft genome assembly for the mermithid nematode *R. culicivora*x was generated from 26.9 gigabases (Gb) of filtered raw data (from a total of 41 Gb sequenced; Table 1). The assembly has a contig span of 267 million base pairs (Mb) and a scaffold span of 323 Mb. The 52 Mb of spanned gaps are likely inflated estimates derived from our use of the SSPACE scaffold. We do not currently have a validated independent estimate of genome size for *R. culicivora*x, but preliminary measurements with Feulgen densitometry suggest a size greater than 320 Mb (Elizabeth Martínez Salazar pers. comm.). The *R. culicivora*x genome is thus likely to be three fold bigger than that of *C. elegans*, and five fold that of *T. spiralis* (Table 2). The assembly is currently in 62,537 scaffolds and contigs larger than 500 bp, with an N50 of 17.6 kb. The N50 for scaffolds larger than 10 kb is 29.9 kb, and the largest scaffold is over 200 kb. The GC content is 36.3%, comparable

to 38% of *C. elegans* and 34% in *T. spiralis*. We identified 47% of the *R. culicivora* genome as repetitive. To validate this estimate we repeated our repeat-finding approach against previously published genomes and achieved good accordance with published data (Table 2). The non-repetitive content of the *R. culicivora* genome is thus approximately twice that of *C. elegans* and three times that of *T. spiralis*. *T. spiralis* thus stands out as having the least complex nematode genome sequenced thus far, and the contrast with *R. culicivora* shows that small genomes are not a characteristic of Dorylaimia.

The RNA-Seq data were assembled into 29,095 isotigs in 22,418 isogroups spanning 23 Mb, and thus are likely to be a reasonable estimate of the *R. culicivora* transcriptome. Using BLAT [36], 21,204 of the isotigs were found to be present (with matches covering >80% of the isotig) in single contigs or scaffolds of the genome assembly, suggesting reasonable biological completeness and contiguity. We also used the CEGMA approach to assess quality of the genome assembly, and found high representation (89.92% partial, 75.40% complete) and low proportion of duplicates (1.05 fold), suggesting a high quality assembly with limited retained haploid assembly duplicates (Table 1). Automated gene prediction from the assembly with iterative rounds of the MAKER pipeline, using the RNA-Seq data as evidence both directly and through GenomeThreader-derived mapping, yielded a total of nearly 50,000 gene models. These were reduced to 48,171 gene models by merging those with identities >99% using Cd-hit. This gene count would be surprisingly high for a nematode: *C. elegans* has ~22,000 genes, *T. spiralis* has ~16,000, and *Pristionchus pacificus* has ~27,000. The excess of *R. culicivora* gene models may result from poorly assembled contigs, from assembly fragmentation, and "over-enthusiastic" prediction from gene modelers within the MAKER pipeline. Within the 48,171 predictions, 12,026 were derived from the Augustus modeler and 36,145 from SNAP. Because Augustus predictions conservatively require some external evidence (transcript mapping and/or sequence similarity to other known proteins), we regarded these as the most reliable and biologically complete. Exons of the Augustus-predicted genes in *R. culicivora* had a median length of 161 bp, slightly larger than those in *C. elegans* (137bp) and *T. spiralis* (128bp). Introns of the *R. culicivora* Augustus models, with a median of 405 bp, were much larger than those in *C. elegans* (69 bp) or *T. spiralis* (283bp). The small introns observed in *C. elegans* and other rhabditid nematodes (Table 2) are thus likely to be a derived feature.

We annotated 1,443 tRNAs in the *R. culicivora* genome using INFERNAL [37] and tRNAscan-SE [38], of which 382 were pseudogenes (see Table S5 for details). In comparison, *T. spiralis* has 134 tRNAs of which 7 are pseudogenes, while *C. elegans* has 606 tRNAs with 36 pseudogenes [39]. Threonine (Thr) tRNAs were particularly overrepresented (676 copies), a finding echoed in the genomes of *Meloidogyne incognita* and *Meloidogyne floridensis* (tylenchine nematodes, see Figure 1) [24] and in *P. pacificus* [20]. *P. pacificus* also

has an overrepresentation of Arginine tRNAs [39].

We have made the annotated *R. culicivora* genome, with functional categorisations of predicted genes and proteins and annotation features, available in a dedicated genome browser at <http://romanomermis.nematod.es>.

The *R.culicivora* proteome retains conserved metazoan components lost in *T. spiralis* and *C.elegans*

The phylogenetic placement of *R. culicivora* compared to *C. elegans* makes its genome ideal for exploring the likely genetic complexity of the ancestral nematode. With *T. spiralis*, it can be used to reveal the idiosyncracies of the several genomes available for Rhabditida. To polarise this comparison, we used data from the genome of the arthropod *T. castaneum*. The *T. castaneum* genome is of high quality [40] and the pattern of development of this beetle is less derived than that of the major arthropod model *Drosophila melanogaster* [41]. We used the orthoMCL pipeline to generate a set of gene clusters for the four species *R. culicivora*, *T. spiralis*, *C. elegans* and *T. castaneum*. The large sequence divergence between the four species may have obscured orthology relationships, making inference of true functional orthology problematic [42–44], but the parameters used (a BLAST E-value of $1e^{-5}$, and orthoMCL inflation parameter of 1.5) can be regarded as relaxed (i.e. most inclusive) compared to other studies [44–46]. As the *R. culicivora* genome assembly may not be complete, we based inference of absence on shared loss in both *R. culicivora* and *T. spiralis*. Thus, we believe that our analyses were at a minimum able to identify homologues where present, and thus we could robustly infer absence. While the orthoMCL pipeline is regarded as very robust in accurately clustering unknown proteins [47] inferences of functional or biological orthology are complex. Inferences of absence were explored in detail (Supplementary file 5).

We identified 3274 clusters that contained protein representatives from all three nematode genomes, and 2833 of these also contained at least one *T. castaneum* representative (Figure 2). These 2833 clusters represent a conserved metazoan and eukaryotic core proteome. There were many clusters that contained proteins from only one species of nematode, representing lineage specific expansions of novel protein families. *T. spiralis* had the lowest number of these (975), while *C. elegans* and *R. culicivora* each had over two thousand. Interestingly, of the 2747 *R. culicivora*-limited clusters, 324 (11.8%) had apparent orthologues in *T. castaneum*. Such clusters are candidates for retention of phylogenetically ancient genes by one nematode species and loss in the other two.

T. spiralis appeared to have lost more phylogenetically ancient genes than had either *R. culicivora* or *C.*

elegans. *T. spiralis* and *C. elegans* shared only 412 clusters exclusive of *R. culicivorax* members, while *R. culicivorax* and *C. elegans* shared 1298 clusters exclusive of *T. spiralis*. Despite their phylogenetic affinity, *R. culicivorax* and *T. spiralis* only shared 600 clusters exclusive of *C. elegans*. *C. elegans* and *R. culicivorax* shared very similar numbers of clusters with *T. castaneum* (2833 contain all species in the comparison; 853 contain only *C. elegans*, *R. culicivorax* and *T. castaneum*, 569 *C. elegans* and *T. castaneum*, and 568 *R. culicivorax* and *T. castaneum*) (Figure 2).

The clusters containing only *R. culicivorax* and *T. spiralis* might identify functions important to these dorylaim nematodes. In the 461 *T. spiralis* and 806 *R. culicivorax* proteins in these clusters, a total of 65 GO terms were found to be overrepresented ($p < 0.05$ by Fisher's exact test) compared with the GO annotation set derived from the complete *C. elegans* proteome, and 33 were overrepresented when compared to annotation of the *T. castaneum* genome. There were 26 GO terms overrepresented in both comparisons. Clusters with *R. culicivorax*, *T. spiralis* and *T. castaneum* members (but lacking *C. elegans* members) contained 332 *R. culicivorax* and 573 *T. spiralis* and 445 *T. castaneum* proteins, and we identified 40 GO terms overrepresented compared to the GO annotated *C. elegans* proteome (see Supplementary file 2). From this we suggest that *T. spiralis* may not have a typical dorylaim genome. The *T. spiralis* genome is reduced in content compared to other nematodes: it is smaller, has fewer genes overall, and has fewer phylogenetically ancient genes. This is congruent with the previously reported loss of proteins with metabolic function in *T. spiralis* [26]. The evolutionary reasons behind this reduction remain obscure, but could include loss of genetic capacity following acquisition of a unique lifestyle that lacks a freeliving stage or genomic streamlining to permit rapid reproduction and growth. Many parasitic and endosymbiotic prokaryotes and eukaryotes have reduced genome sizes [48].

The genetic background of development in *R. culicivorax* and *T. spiralis* differs markedly from that of *C. elegans*

In a recent multi-species developmental timecourse expression analysis within the genus *Caenorhabditis*, conserved sets of genes were found to be over-expressed in discrete portions of the developmental timeline from zygote to hatching larva [49]. In particular, this study suggests a conserved period in development where a very restricted set of genes is expressed in all species, perhaps corresponding to a 'bauplan' stage in nematode development as has been proposed for Metazoa in general. To explore whether this model

can be extended across Nematoda, we identified *R. culicivora*x and *T. spiralis* homologues of the 1725 developmentally regulated *C. elegans* genes extracted from this analysis [49]. Nearly half (845) of these genes were not grouped in clusters with Dorylaimia proteins using orthoMCL. We were unable to identify any sequence homologs for 450 of the proteins in *R. culicivora*x using BLAST+.

The remaining 395 proteins had BLAST+ hits to *R. culicivora*x proteins, but were so divergent that orthoMCL did not cluster them as orthologs with Dorylaimia proteins. Among these 395 with marginal matches, we found that 18 belonged to the *C. elegans* nuclear hormone receptor subfamilies, 5 were innexin type gap-junction protein, 6 were TWiK potassium channel proteins and 5 were acetylcholine receptor proteins. These protein families are particularly diverse and expanded in *C. elegans* [50–53] and we suggest that the genes "missing" from *R. culicivora*x but having low-scoring BLAST+ matches represent rapidly evolved, divergent duplications within the lineage leading to *C. elegans*. OrthoMCL is likely to be correct in not clustering most of these proteins. The proportion of *Caenorhabditis*-restricted genes across the developmental timecourse examined by Levin et al. [49] varied from 36.4% to 59.9% (Figure 3 and Supplementary file 4). A surprisingly high proportion of the developmental genes acting during specific embryonic stage transitions appear to be unique to the genus *Caenorhabditis* or at least so divergent that functional orthology, including interaction with conserved partners, is doubtful. A striking difference between *R. culicivora*x and *T. spiralis* was apparent, with 238 of the developmentally differentially expressed *C. elegans* genes having a *R. culicivora*x homologue but not a *T. spiralis* homologue, while only 88 had a *T. spiralis* homologue but not an *R. culicivora*x one. Given the conservatism of body plan evolution in nematodes, these dramatic genetic differences suggest extensive, largely phenotypically "silent" changes in the genetic programmes orchestrating nematode development. We used computational comparisons of selected key molecular processes and pathways to tease out the differences between the model *C. elegans* and the two dorylaim species, *T. spiralis* and *R. culicivora*x.

Core developmental pathways differ between nematodes

There are important differences in the cellular biology of development between *R. culicivora*x and *C. elegans* [34, 35], and we used the genomic data to follow up on some of the more striking contrasts between the dorylaim and the rhabditid patterns of development: primary axis polarity, segregation of maternal message within the early embryo, hypodermis formation, the vulval specification pathways, epigenetic pathways (especially DNA methylation), sex determination and light sensing.

In the *C. elegans* 2-cell stage mitotic spindles rotate 90° in the posterior germline cell, and the subsequent cell divisions are orthogonal [54–56]. This rotation is not observed in *R. culicivora* and division is longitudinal [34]. In *C. elegans* and many other animals *par* genes are essential for cell polarisation [57] and polarised distribution of PAR proteins results in the restriction of mitotic spindle rotation to one cell. *C. elegans* mutants lacking *par-2* and *par-3* genes resemble the *R. culicivora* phenotype, showing longitudinal spindle orientation [58]. The *par-2* gene was missing from both *R. culicivora* and *T. spiralis* (Figure 3; Table 3). Additionally, no orthologues for the *par-2*-interacting genes *let-99*, *gpr-1* or *gpr-2*, required for proper embryonic spindle orientation in *C. elegans* [59], were identified in the dorylaims using orthoMCL clustering or sensitive BLAST searches. We identified a candidate *par-3* in *R. culicivora*, but this was so divergent from *C. elegans*, *T. castaneum* and *T. spiralis par-3* that these putative orthologues were not clustered in our analysis. The *D. melanogaster par-3* ortholog *bazooka* functions in anterior-posterior axis formation, but as in *R. culicivora* and *T. spiralis par-2* is absent from the fly [60]. Thus, we hypothesise that the PAR-3 - PAR-2 system for regulating spindle positioning evolved within in the lineage leading to the genus *Caenorhabditis*. The divergent *par-3*-like gene in dorylaims may be involved in axis formation, but perhaps interacts with different partner proteins.

Once polarity has been established in the early *C. elegans* embryo, many maternal messages are differentially segregated into anterior or posterior blastomeres [56,61]. MEX-3 is an RNA-binding protein translated from maternally-provisioned mRNAs found predominantly in early anterior blastomeres [62,63]. We identified a highly divergent MEX-3 homologue in *R. culicivora*, but found no orthologue in *T. spiralis*.

To demonstrate the utility of the *R. culicivora* system, and the power of the genome-to-development model, we assayed its expression in embryos using *in situ* hybridisation. We selected the *mex-3* gene for these studies, as it is strongly expressed and highly localised during a short time window in development in *C. elegans*. The observed expression pattern in *R. culicivora* is similar to *C. elegans* (Figure 5). In the fertilized *R. culicivora* egg *mex-3* mRNA is initially equally distributed. Prior to first cleavage *mex-3* mRNA is segregated to the anterior pole and thus becomes essentially restricted to the somatic S1 blastomere (for nomenclature, see [14]). With the division of S1 it is localised to both daughter cells. After the 4-cell stage the signal disappears gradually. Despite the presence, and apparent conservation of expression pattern, of *mex-3*, we were unable to identify other components of the *C. elegans* maternal mRNA regulation system, such as *mex-5*, *mex-6* and *spn-4* in either dorylaim species. While MEX-5 and MEX-6 are important for controlled MEX-3 expression in *C. elegans* [64], the apparent absence of SPN-4 in *R. culicivora* and *T. spiralis* is particularly intriguing. SPN-4 links embryonic polarity conferred by the *par* genes and partners

to cell fate specification through maternally deposited mRNAs and proteins [65,66]. This suggests that the core regulatory logic of the early control of axis formation and cell fate specification must differ significantly between the dorylaim species and *C. elegans*.

The hypodermis in *C. elegans* is derived from specific descendants of the anterior S1 (AB) and the posterior S3 (C) founder cells [67]. In contrast, in *R. culicivora* the hypodermis is derived from S2 (EMS) descendants, which form repetitive ring structures that extend from posterior to anterior [35]. Several developmental regulatory genes expressed in the hypodermis or associated with hypodermal development were present only in *C. elegans* in our analysis (see Table 3 and Supplementary file 4). The GATA-like transcription factor gene *elt-3* gene is absent in the dorylaim species, but the *elt-3* ortholog *elt-1* is conserved in *R. culicivora*, *T. spiralis* and *T. castaneum*. These genes act redundantly in *C. elegans* hypodermis formation [68]. Thus *elt-3* involvement must be an innovation in the rhabditid lineage, suggesting changes of interaction complexity during nematode evolution.

In *C. elegans*, vulva formation is highly dependent on the initial cell-cell interactions of the anchor cell with the neighboring vulva precursor cells (VPCs). Induction of the VPCs activates a complex gene regulatory network which drives divisions and differentiations of the VPCs to form a functional vulva. The evolutionary lability of this system has been explored in rhabditid nematodes, revealing the changing relative importances of a series of cell-cell interactions, short and long range inductions, and lineage-autonomous specifications [69,70]. The signal transduction pathways involved include a RTK/RAS/MAPK cascade, activated by EGF- and wnt-signaling [71]. Among the downstream targets in *C. elegans* are for example *lin-1* and the β -catenin *bar-1*, which in turn regulates the HOX-5 ortholog *lin-39* [72–74]. Our analysis shows that *lin-1* and *bar-1*, as well as other important regulators of vulva development, are absent from the genomes of *R. culicivora* and/or *T. spiralis* (Table 3 and Supplementary File 4). We identified a *R. culicivora* gene with a low-quality match to BAR-1 (24.2% sequence identity). This protein is not clustered with other dorylaim proteins, and appears to be either a duplication of the β -catenin ortholog HMP-2 or another armadillo repeat-containing protein and not orthologous to *bar-1* (see Supplementary file 5). These shared patterns of absence again indicate that the same morphological structures can be generated with very different genetic underpinnings. While it is possible that vulva formation in the dorylaims is regulated without the *bar-1* - *lin-39* interactions, as observed in *P. pacificus* [75], it may be that HOX genes function differently in the dorylaims: rather than acting in a lineage-dependent manner (as in *C. elegans* [76,77]) they may act in a positional regulatory manner, as in other animals [78,79].

Epigenetic regulation is key to developmental processes in many animals, but its roles in *C. elegans* are more

mutated. While *C. elegans* has a reduced ability to methylate DNA [80], orthologue clusters restricted to *R. culicivorax* and *T. spiralis* (excluding *C. elegans*) were enriched for four methylation-associated GO terms. We also found significant enrichment ($p < 0.05$) for GO terms describing chromatin and DNA methylation functions in the set of *R. culicivorax* proteins that lacked homologues in *C. elegans* (see Supplementary file 2). Important roles for methylation and changes in methylation patterns in the development of *T. spiralis* have been inferred from transcriptional profiling [81]. In addition, methylation is important for the silencing of transposable elements [82,83] and could play a crucial role in the highly repetitive *R. culicivorax* genome. The *C. elegans* genome was also found to be depleted for chromatin re-modeling genes of the Polycomb and Trithorax groups [84]. It is intriguing that we found orthologs of *T. castaneum pleiohomeotic* in *R. culicivorax* and *T. spiralis*, and orthologs of *T. castaneum trithorax* and *Sex comb on midleg (Scm)* in *R. culicivorax*. This suggests that dorylaim chromatin restructuring mechanisms may be much more arthropod-like than are those of *C. elegans*. The presence of an intact methylation machinery and conserved chromatin re-modelling factors opens the prospects for a role for epigenetic modification in developmental regulation in dorylaim nematodes.

Sex determination machinery

The mechanism of sex determination differs considerably among animals and it has been claimed to be one of the developmental programs most influenced by developmental system drift [16]. Sex ratios in *R. culicivorax* are described to be environmentally determined through in-host nematode density [85], and thus might be fundamentally different from the system found and extensively analysed in *C. elegans* [86]. Environmental sex determination is found in many nematode taxa, including Strongyloididae and Meloidogyninae (both Tylenchina), taxa more closely related to *C. elegans*. *C. elegans* sex determination is based on the X to autosome ratio, with males haploid for the X chromosome (XO), and females diploid (XX). This difference is read by the master switch *xol-1* [87], which acts through the three *sdc* genes [88–90] to regulate the systemic secretion of HER-1, a ligand for the TRA-2 receptor [91–93]. TRA-2 in turn negatively regulates a complex of *fem* genes, which regulates nuclear translocation of TRA-1, the final shared step in the pathway that switches between male and female systems. We did not find credible homologues (through orthoMCL and re-confirmation with BLAST) of *xol-1*, *sdc-1*, *sdc-2*, *sdc-3*, *her-1* or *tra-2* in either *T. spiralis* or *R. culicivorax* (Table 3; Supplementary file 5), and thus these species are unlikely to use the HER-1/TRA-2 ligand-receptor system to coordinate organism-wide sexual differentiation.

Light sensing machinery in *R. culicivora*

Light sensing with and without eye-like organs has been described in other mermithids [94,95]. Although *R. culicivora* has no structurally evident eye spots it is likely that invasion of the mosquito host on the surface of the water body [32] and migration of emerged nematodes migrate back to the substrate to mate and deposit eggs [96] involves phototactic behaviour. Preliminary experiments with *R. culicivora* give support to this view (J. Burr, pers. comm.), but the underlying physiology has not been explored. We identified several GO terms associated with photoreceptor development and light sensing (see Supplementary file 2) in *R. culicivora* proteins in comparison to *C. elegans* and *T. castaneum* proteomes (in the set of *R. culicivora* proteins without homologues in these species). Two especially intriguing GO terms were 'phototaxis' and 'energy taxis'. Proteins associated with these GO terms had BLAST similarities to COUP transcription factors, which in the mouse have been associated with cell fate determination in the eye [97].

In *Mermis nigrescens*, a close relative of *R. culicivora*, a directional light sensing organ is found in the anterior pharynx, where a cylinder of light-shadowing cells packed with a nematode hemoglobin shades a central photoreceptor [94,98,99]. Globin-like domains are found in diverse gene families in Nematoda [100]. More *R. culicivora* proteins were annotated with the GO term 'oxygen binding' than those of the other species analysed (Supplementary figure 1). Several of these *R. culicivora* proteins have BLAST matches to bonafide globins and hemoglobins, and an optical shadowing function is possible for one or more of them. Pigment granules are segregated into the hypodermis of *R. culicivora* (see Figure 5) and may also have a light-shadowing function [34]. We also found the GO term 'cellular pigment accumulation' in the set of *R. culicivora* proteins that had homologues with *T. spiralis* and *T. castaneum*, but not with *C. elegans*. The protein associated with this GO term was most similar to *Xenopus* SHROOM2 protein, which is involved in melanosome formation and expressed in the eye of the frog [101]. We also identified a candidate opsin in *R. culicivora*. The gene is partially supported by EST data, and could generate a 313 amino acid protein with identities of 26% to the *Bos taurus* (accession NP_776991) and *Didelphis aurita* (ABC75817) long-wave-sensitive opsins.

Conclusions

By combining the *R. culicivora* genome presented here together with the published *T. spiralis* genome, we have been able to explore the molecular diversity of Dorylaimia, and provide robust contrasts with the intensively studied Rhabditida. Particularly surprising were the differences between *R. culicivora* and *T. spiralis*. The *R. culicivora* genome is much larger than that of *T. spiralis*. A majority of the genome

was identified as repetitive, including many transposable elements. Despite the phylogenetic and lifestyle affinities between the two dorylaims compared to *C. elegans*, the *R. culicivora* genome retained many more genes in common with *C. elegans* than did *T. spiralis*. We suggest that *T. spiralis* may be an atypical representative of dorylaim nematodes, perhaps due to a highly derived life cycle.

Our analyses identified many genes apparently absent from the dorylaim genomes. We used very relaxed analysis parameters, and performed close analyses of genes identified as critical in *C. elegans* development for which we could find no credible dorylaim orthologues. In these phylum-spanning comparisons, inferences of gene orthology can be obscured by levels of divergence. In addition, the gene family birth rate in the chromadorean lineage leading to *C. elegans* is high [25,26], and therefore *C. elegans* was expected to have many genes absent from the dorylaim species. Thus, we might not have found a *R. culicivora* orthologue for a specific gene for three reasons: it may have arisen in the branch leading to *C. elegans*; its sequence divergence may be too great to permit clustering with potential homologs; or it was not assembled in the draft dorylaim genomes. The analyses of *C. elegans* PAR-3 and *D. melanogaster* *bazooka* illustrate some of these difficulties: the possible *R. culicivora* orthologue was highly divergent. Whether or not we have been able to identify all the orthologues of the key *C. elegans* genes present in the *R. culicivora* and *T. spiralis* genomes, the absence of an identified orthologue maximally implies loss from the genome, and minimally implies significant sequence, and thus functional, divergence.

Between the model organisms *C. elegans* and *D. melanogaster* many key mechanisms governing early cell patterning are divergent [54]. Our data indicate that a major divergence also exists within Nematoda. *T. spiralis* and *R. culicivora* share a lack of orthologues of genes involved in several core developmental processes in *C. elegans*, and many of these *C. elegans* genes are restricted to the Rhabditida. It is thus doubtful that these processes are regulated by same molecular interactions across the phylum. To the contrary it is likely that developmental system drift has played (and still plays) a major role in nematode evolution. The phenotypic conservatism associated with the vermiform morphology of nematodes [102] has fostered unjustified expectations concerning the genetic programmes that determine these morphologies.

To be useful as a contrasting system to the 'canonical' *C. elegans* model, any nematode species must be accessible to both descriptive and manipulative investigation. Here, we have defined a reference genome for *R. culicivora*, laying bare the core machinery available for developmental regulation, and demonstrated that *in situ* hybridisation approaches are feasible for this species. Along with the robust laboratory cultures long established, this makes *R. culicivora* an attractive and tractable alternative model for understanding the evolutionary dynamics of nematode developmental biology. We have highlighted a few of the possible avenues

a research programme could follow: early axis formation and polarisation, the specification of hypodermis, sex determination, vulva formation and the roles of epigenetic processes in developmental regulation. The advent of robust, affordable and rapid genome sequencing also opens the vista of large-scale comparative genomics of development across the phylum Nematoda [28] to better understand the diversity of the phylum and also place the remarkable *C. elegans* model in context of its peers. It will next be necessary to extend these analyses to a broader sampling of developmental pathway genes from a wider and fully representative sampling of nematode genomes across the full diversity of the phylum.

Methods

Sequencing and Genome Assembly

Genomic DNA was extracted from several hundred, mixed-sex, adult *R. culicivora* specimens from a culture first established in Ed Platzer's laboratory in Riverside, California. Illumina paired end and mate pair sequencing with libraries of varying insert sizes, and Roche 454 single end sequencing, was performed at the Cologne Center for Genomics - CCG (<http://www.ccg.uni-koeln.de>). A Roche 454 dataset of transcriptome reads from cDNA synthesised from mixed developmental stages and sexes was also generated (see Table S1 for details of data generation).

The quality of the raw data was assessed with FastQC (v.0.9) (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>). Adapter sequences and low quality data were trimmed from the Illumina paired end data with custom scripts (see <http://github.com/sujaikumar/assemblage>) and from the mate pair libraries with Cutadapt (v.1.0) [103]. We constructed a preliminary genome assembly, with relaxed insert size parameters, from the paired end Illumina libraries with the de-novo-assemble option of the clcAssemblyCell (v.4.03b) [104]. We validated the actual insert sizes of our libraries by mapping back the reads to this preliminary assembly using clcAssemblyCell. The preliminary assembly was also used to screen out bacterial and other contaminant data [105]. The transcriptome data were assembled with Roche GSAssembler (Newbler; version 2.5). For the production assembly, we explored assembly parameters using different mixes of our data, evaluating each for total span, maximal contig lengths, N50, number of contigs, representation of the transcriptome, and conserved eukaryotic gene content (using the CEGMA pipeline in version 2.1 [106]). The most promising assembly was scaffolded with the filtered Illumina mate pair read sets using SSPACE (v.1.2) [107]. As our genomic DNA derived from a population of nematodes of unknown genetic diversity, we removed short contigs that mapped entirely within larger ones using Cd-hit (v.4.5.7) [108] at a 95% cutoff. A final round of superscaffolding was performed, linking scaffolds that had logically consistent matches to

the transcriptome data based on BLAT [36] hits and processed with SCUBAT (B. Elsworth, pers. comm.; <http://github.com/elswob/SCUBAT>). The final genome assembly was again assessed for completeness by assessing the mapping of the transcriptome contigs and with the CEGMA pipeline [106].

Genome Annotation

RepeatMasker (v.3.3.0) [109, 110], RepeatFinder [111] and RepeatModeler (v.1.0.5) (<http://www.repeatmasker.org/RepeatModeler.html>; combining RECON (v.1.07) [112] and RepeatScout (v.1.05) [113]), were used to identify known and novel repetitive elements in the *R. culicivora* genome. We employed the MAKER pipeline [114] to find genes in the *R. culicivora* genome assembly. In a first pass, the SNAP gene predictor included in MAKER was trained with a CEGMA [106] derived output of predicted highly conserved genes. As additional evidence we included the transcriptome assembly and a set of approximately 15,000 conserved nematode proteins derived from the NEMBASE4 database [115] (recalculated by J. Parkinson; pers. comm.). In the second, definitive, pass we used the gene set derived from this first MAKER iteration to train Augustus [116] inside the MAKER pipeline for a second run, also including evidence from transcriptome to genome mapping obtained with GenomeThreader [117]. Codon usage in *R. culicivora*, *T. spiralis* and *C. elegans* was calculated using INCA (v2.1) [118]. Results were then compared to data from [119] (see Supplementary files 1 and 3).

We used Blast2GO (Blast2GO4Pipe, v.2.5, January 2012 database issue) [120] to annotate the gene set with Gene Ontology terms [121], based on BLAST matches with expect values less than $1e^{-5}$ to the UniProt/SwissProt database (March 2012 snapshot), and domain annotations derived from the InterPro database [122]. Comparison of annotations between three nematode species (*R. culicivora*, *C. elegans* and *T. spiralis*) and, as a reference outgroup, the holometabolous coleopteran arthropod *Tribolium castaneum* was based on GO Slim data retrieved with Blast2GO. RNA genes were predicted using INFERNAL (v.1.0.2) [37] and the Rfam database [123], and tRNAscan-SE (v.1.3.1) [38].

Orthology Screen

We inferred clusters of orthologous proteins between *R. culicivora*, *T. spiralis* and *C. elegans*, and the beetle *T. castaneum* using OrthoMCL (v.2.0.3) [124]. *T. spiralis*, *C. elegans* and *T. castaneum* protein sets were downloaded from NCBI and WormBase (see Table S2) and redundancy screened with Cd-hit at the 99% threshold. We selected an inflation parameter of 1.5 for MCL clustering (based on [125, 126]) within OrthoMCL to generate an inclusive clusterings in our analysis likely to contain even highly diverged

representatives from the four species. In analyses of selected developmental genes, clusters were manually validated using NCBI-BLAST+ [127]. We affirmed the uniqueness of *C. elegans* proteins identified as lacking homologues in the enoplean nematodes by comparing them to the *R. culicivora* proteome using BLAST. Those with no significant matches at all (all matches with E-values $> 1e^{-5}$) were classified as confirmed absent. Those having matches with E-values $< 1e^{-5}$ were investigated further by surveying the cluster memberships of the *R. culicivora* matches. If the *R. culicivora* protein was found to cluster with a different *C. elegans* protein, the uniqueness to *C. elegans* was again confirmed. If the *R. culicivora* protein did not cluster with an alternative *C. elegans* protein, we reviewed the BLAST statistics (E-value, identity and sequence coverage) of the match and searched the GenBank non redundant protein database for additional evidence of possible orthology. Only if these tests yielded no indication of direct orthology was the *C. elegans* protein designated absent from the enoplean set. Further details of the process are given in Supplementary file 5.

We identified the protein sequences of 1,725 genes differentially expressed in *C. elegans* developmental stages [49] and selected, using our OrthoMCL clustering, those apparently lacking orthologues in *R. culicivora* and *T. spiralis* (verified as above). Using Wormbase (<http://www.wormbase.org>, release WS233) we surveyed the *C. elegans*-restricted genes for their experimentally-defined roles in development.

Custom Perl scripts were used to group orthoMCL clusters on the basis of species membership patterns. The sets of clusters that contained (i) both *T. spiralis* and *R. culicivora* members but no *C. elegans* members and (ii) *T. spiralis* and *R. culicivora* and *T. castaneum* members but no *C. elegans* members were surveyed for GO annotations enriched in comparison to the whole *C. elegans* proteome (sets i and ii) and the *T. castaneum* proteome (set i), conducting Fisher's exact test as implemented in Blast2GO. To improve annotation reliability, these proteins were recompiled (using BLAST) to the UniProt/SwissProt database and run through the Blast2GO pipeline in the same way as described above.

Whole-mount in situ hybridization

For in situ hybridisation we modified the freeze-crack procedure described previously for *C. elegans* [128] and revised by Maduro et al. (2007; <http://www.faculty.ucr.edu/~mmaduro/resources.htm>). In particular to allow for reliable penetration of the durable *R. culicivora* egg envelopes we initially partly removed the protective layer by incubation in alkaline bleach solution (see [34]). Digoxigenine-labeled sense and antisense RNA probes were generated from linearized pBs vectors (Stratagene, La Jolla, USA) containing a 400 bp fragment of *R. culicivora mex-3* via run off *in vitro* transcription with T7 or T3 RNA-polymerase according

to the manufacturer's protocol (Roche, Mannheim, Germany). The concentration of the labeled probes was about $300 \text{ ng} \times \text{ml}^{-1}$.

Author's contributions

PHS conceived study, assembled and annotated the genome, conducted analyses and wrote paper; MK conceived study, conducted analyses and wrote part of the paper; CK conceived part of the study, conducted analyses on developmental expression set and wrote part of the paper; GDK helped with genome assembly and annotation; SK helped with genome assembly and wrote/provided Perl scripts; JIRC analysed MEX-3 dataset; NAN analysed PAR dataset; DS analysed SEX determination dataset; KM conducted RNA sequencing and initial EST assembly; PH performed preparative laboratory experiments and conceived sequencing strategy; JA conceived sequencing strategy and conducted genome sequencing; PF helped with initial genome pre-assembly; PN initiated study and conceived sequencing strategy; WKT conceived parts of study; MLB conceived study and wrote paper; ES initiated and conceived study and wrote paper

Funding

This work was partly funded through the SFB 680; "Molecular Basis of Evolutionary Innovations". Philipp H. Schiffer is funded by the VolkswagenStiftung in the "Förderinitiative Evolutionsbiologie". Gerogios D. Koutsovoulos is funded by a UK BBSRC Research Studentship and an Overseas Reasearch Studentship from the University of Edinburgh.

Acknowledgements

We are indebted to E. Platzer, Riverside, for the continuous supply with *R. culicivora* nematodes. We thank J. Schulze, Cologne, for advice on nematode cultivation and C. Becker and K. Konrad for expert technical assistance in the genome sequencing experiments. We are also grateful to H. Oezden, Cologne for assistance with In-situ hybridisations. We thank J. Parkinson, Toronto, for providing a conserved NEMBASE4 protein set, Elizabeth Martínez Salazar, Zacatecas, Mexico, for Feulgen C-value data and J. Burr, Vancouver, Canada for sharing preliminary results on phototaxis in *R. culicivora*.

Assemblies and other computations were conducted on the HPC cluster "CHEOPS" at the University of Cologne (<http://rrzk.uni-koeln.de/cheops.html>).

References

1. Blaxter M, de Ley P, Garey J, Liu L, Scheldeman P, Vierstraete A, Vanfleteren J, Mackey L, Dorris M, Frisse L: **A molecular evolutionary framework for the phylum Nematoda.** *Nature* 1998, **392**(6671):71–75.
2. Maduro MF: **Cell fate specification in the *C. elegans* embryo.** *Developmental Dynamics* 2010, **239**:1315–1329.
3. Sulston JE, Schierenberg E, White JG, Thomson JN: **The embryonic cell lineage of the nematode *Caenorhabditis elegans*.** *Developmental Biology* 1983, **100**:64–119.
4. Boveri T: **Die Entwicklung von *Ascaris megalocephala* mit besonderer Ruecksicht auf die Kern-verhaeltnisse.** *Festschrift fuer C. von Kupffer. Jena: Fischer* 1899, :383–430.
5. Müller H: **Beitrag zur Embryonalentwicklung von *Ascaris megalocephala*.** *Zoologica* 1903, **41**.
6. Vangestel S, Houthoofd W, Bert W, Borgonie G: **The early embryonic development of the satellite organism *Pristionchus pacificus*: differences and similarities with *Caenorhabditis elegans*.** *Nematology* 2008, **10**:301–312.
7. Skiba F, Schierenberg E: **Cell lineages, developmental timing, and spatial pattern formation in embryos of free-living soil nematodes.** *Developmental Biology* 1992, **151**(2):597–610.
8. Lahl V, Schulze J, Schierenberg E: **Differences in embryonic pattern formation between *Caenorhabditis elegans* and its close parthenogenetic relative *Diploscapter coronatus*.** *The International journal of developmental biology* 2009, **53**(4):507–515.
9. Brauchle M, Kiontke K, Macmenamin P, Fitch DHA, Piano F: **Evolution of early embryogenesis in rhabditid nematodes.** *Developmental Biology* 2009, **335**:253–262.
10. Wiegner O, Schierenberg E: **Specification of Gut Cell Fate Differs Significantly between the Nematodes *Acroboloides nanus* and *Caenorhabditis elegans*.** *Developmental Biology* 1998, **204**:3–14.
11. Wiegner O, Schierenberg E: **Regulative development in a nematode embryo: a hierarchy of cell fate transformations.** *Developmental Biology* 1999, **215**:1–12.
12. Meldal B, Debenham N, de Ley P, de Ley I, Vanfleteren J, Vierstraete A, Bert W, Borgonie G, Moens T, Tyler P, Austen M, Blaxter M, Rodgers A, Lambshhead P: **An improved molecular phylogeny of the Nematoda with special emphasis on marine taxa.** *Molecular Phylogenetics And Evolution* 2007, **42**(3):622–636.
13. Voronov DA, Panchin YV: **Cell lineage in marine nematode *Enoplus brevis*.** *Development* 1998, **125**:143–150.
14. Schulze J, Schierenberg E: **Evolution of embryonic development in nematodes.** *EvoDevo* 2011, **2**:18.
15. Schulze J, Houthoofd W, Uenk J, Vangestel S, Schierenberg E: **Plectus - a stepping stone in embryonic cell lineage evolution of nematodes.** *EvoDevo* 2012, **3**:13.
16. True JR, Haag ES: **Developmental system drift and flexibility in evolutionary trajectories.** *Evolution and Development* 2001, **3**(2):109–119.
17. C elegans Sequencing Consortium: **Genome sequence of the nematode *C. elegans*: a platform for investigating biology.** *Science* 1998, **282**(5396):2012–2018.
18. Stein LD, Bao Z, Blasiar D, Blumenthal T, Brent MR, Chen N, Chinwalla A, Clarke L, Clee C, Coghlan A, Coulson A, D'eustachio P, Fitch DHA, Fulton LA, Fulton RE, Griffiths-Jones S, Harris TW, Hillier LW, Kamath R, Kuwabara PE, Mardis ER, Marra MA, Miner TL, Minx P, Mullikin JC, Plumb RW, Rogers J, Schein JE, Sohrmann M, Spieth J, Stajich JE, Wei C, Willey D, Wilson RK, Durbin R, Waterston RH: **The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics.** *PLoS Biology* 2003, **1**(2):E45.
19. Mortazavi A, Schwarz EM, Williams B, Schaeffer L, Antoshechkin I, Wold BJ, Sternberg PW: **Scaffolding a *Caenorhabditis* nematode genome with RNA-seq.** *Genome Research* 2010, **20**(12):1740–1747.
20. Dieterich C, Clifton SW, Schuster LN, Chinwalla A, Delehaunty K, Dinkelacker I, Fulton L, Fulton R, Godfrey J, Minx P, Mitreva M, Roeseler W, Tian H, Witte H, Yang SP, Wilson RK, Sommer RJ: **The *Pristionchus pacificus* genome provides a unique perspective on nematode lifestyle and parasitism.** *Nature Genetics* 2008, **40**(10):1193–1198.

21. Jex AR, Liu S, Li B, Young ND, Hall RS, Li Y, Yang L, Zeng N, Xu X, Xiong Z, Chen F, Wu X, Zhang G, Fang X, Kang Y, Anderson GA, Harris TW, Campbell BE, Vlaminck J, Wang T, Cantacessi C, Schwarz EM, Ranganathan S, Geldhof P, Nejsun P, Sternberg PW, Yang H, Wang J, Wang J, Gasser RB: **Ascaris suum draft genome**. *Nature* 2011, **479**(7374):529–533.
22. Ghedin E, Wang S, Spiro D, Caler E, Zhao Q, Crabtree J, Allen JE, Delcher AL, Guiliano DB, Miranda-Saavedra D, Angiuoli SV, Creasy T, Amedeo P, Haas B, El-Sayed NM, Wortman JR, Feldblyum T, Tallon L, Schatz M, Shumway M, Koo H, Salzberg SL, Schobel S, Pertea M, Pop M, White O, Barton GJ, Carlow CKS, Crawford MJ, Daub J, Dimmic MW, Estes CF, Foster JM, Ganatra M, Gregory WF, Johnson NM, Jin J, Komuniecki R, Korf I, Kumar S, Laney S, Li BW, Li W, Lindblom TH, Lustigman S, Ma D, Maina CV, Martin DMA, McCarter JP, McReynolds L, Mitreva M, Nutman TB, Parkinson J, Peregrin-Alvarez JM, Poole C, Ren Q, Saunders L, Sluder AE, Smith K, Stanke M, Unnasch TR, Ware J, Wei AD, Weil G, Williams DJ, Zhang Y, Williams SA, Fraser-Liggett C, Slatko B, Blaxter ML, Scott AL: **Draft genome of the filarial nematode parasite Brugia malayi**. *Science* 2007, **317**(5845):1756–1760.
23. Godel C, Kumar S, Koutsovoulos G, Ludin P, Nilsson D, Comandatore F, Wrobel N, Thompson M, Schmid CD, Goto S, Bringaud F, Wolstenholme A, Bandi C, Epe C, Kaminsky R, Blaxter M, Mäser P: **The genome of the heartworm, Dirofilaria immitis, reveals drug and vaccine targets**. *The FASEB Journal* 2012, **26**.
24. Abad P, Gouzy J, Aury JM, Castagnone-Sereno P, Danchin EGJ, Deleury E, Perfus-Barbeoch L, Anthouard V, Artiguenave F, Blok VC, Caillaud MC, Coutinho PM, Dasilva C, De Luca F, Deau F, Esquibet M, Flutre T, Goldstone JV, Hamamouch N, Hewezi T, Jaillon O, Jubin C, Leonetti P, Magliano M, Maier TR, Markov GV, McVeigh P, Pesole G, Poulain J, Robinson-Rechavi M, Sallet E, Segurens B, Steinbach D, Tytgat T, Ugarte E, van Ghelder C, Veronico P, Baum TJ, Blaxter M, Bleve-Zacheo T, Davis EL, Ewbank JJ, Favery B, Grenier E, Henrissat B, Jones JT, Laudet V, Maule AG, Quesneville H, Rosso MN, Schiex T, Smant G, Weissenbach J, Wincker P: **Genome sequence of the metazoan plant-parasitic nematode Meloidogyne incognita**. *Nature Biotechnology* 2008, **26**(8):909–915.
25. Kikuchi T, Cotton JA, Dalzell JJ, Hasegawa K, Kanzaki N, McVeigh P, Takanashi T, Tsai IJ, Assefa SA, Cock PJA, Otto TD, Hunt M, Reid AJ, Sanchez-Flores A, Tsuchihara K, Yokoi T, Larsson MC, Miwa J, Maule AG, Sahashi N, Jones JT, Berriman M: **Genomic Insights into the Origin of Parasitism in the Emerging Plant Pathogen Bursaphelenchus xylophilus**. *PLoS Pathogens* 2011, **7**(9):e1002219.
26. Mitreva M, Jasmer DP, Zarlenga DS, Wang Z, Abubucker S, Martin J, Taylor CM, Yin Y, Fulton LA, Minx P, Yang SP, Warren WC, Fulton RS, Bhonagiri V, Zhang X, Hallsworth-Pepin K, Clifton SW, McCarter JP, Appleton J, Mardis ER, Wilson RK: **The draft genome of the parasitic nematode Trichinella spiralis**. *Nature Genetics* 2011, **43**(3):228–235.
27. Hope IA: **Embryology, Developmental Biology and the Genome**. In *The Biology of Nematodes*, Edited by Lee, D. L., New York: Taylor & Francis. 1st edition :2002:121–145.
28. Kumar S, Schiffer PH, Blaxter M: **959 Nematode Genomes: a semantic wiki for coordinating sequencing projects**. *Nucleic Acids Research* 2011, :1–6.
29. Kumar S, Koutsovoulos G, Kaur G, Blaxter M: **Toward 959 nematode genomes**. *Worm* 2012, **1**:0–8.
30. Petersen JJ: **Nematodes as biological control agents: Part I. Mermithidae**. In *Advances in Parasitology*, Edited by Baker, J.R. and Muller, R., London, Academic Press :1985:307–346.
31. Petersen JJ, Chapman HC, Willis OR, Fukuda T: **Release of Romanomermis culicivorax for the control of Anopheles albimanus in El Salvador II. Application of the nematode**. *The American journal of tropical medicine and hygiene* 1978, **27**(6):1268–1273.
32. Shamseldeen MM, Platzer EG: **Romanomermis culicivorax: Penetration of larval mosquitoes**. *Journal of Invertebrate Pathology* 1989, **54**(2):191–199.
33. Mitreva M, Jasmer D: **Biology and genome of Trichinella spiralis**. In *WormBook*, Edited by The C. elegans Research Community, WormBook 2006 .
34. Schulze J, Schierenberg E: **Cellular pattern formation, establishment of polarity and segregation of colored cytoplasm in embryos of the nematode Romanomermis culicivorax**. *Developmental Biology* 2008, **315**(2):426–436.

35. Schulze J, Schierenberg E: **Embryogenesis of *Romanomermis culicivorax*: an alternative way to construct a nematode.** *Developmental Biology* 2009, **334**:10–21.
36. Kent W: **BLAT—the BLAST-like alignment tool.** *Genome Research* 2002, **12**(4):656.
37. Nawrocki EP, Kolbe DL, Eddy SR: **Infernal 1.0: inference of RNA alignments.** *Bioinformatics* 2009, **25**(10):1335–1337.
38. Lowe TM, Eddy SR: **tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence.** *Nucleic Acids Research* 1997, **25**(5):0955–0964.
39. Kumar S: **Data for PhD Thesis on Next Generation Nematode Genomes.** <http://dx.doi.org/10.6084/m9.figshare.96089> 2012.
40. Richards S, and The *T. castaneum* Genome Consortium: **The genome of the model beetle and pest *Tribolium castaneum*.** *Nature* 2008, **452**(7190):949–955.
41. Schröder R, Beermann A, Wittkopp N, Lutz R: **From development to biodiversity—*Tribolium castaneum*, an insect model organism for short germband development.** *Development Genes and Evolution* 2008, **218**(3-4):119–126.
42. Jensen RA: **Orthologs and paralogs - we need to get it right.** *Genome Biology* 2001, **2**(8):interactions1002.1—1002.3.
43. Koonin E: **Orthologs, paralogs, and evolutionary genomics.** *Annual Review of Genetics* 2005, **39**:309–38.
44. Moreno-Hagelsieb G, Latimer K: **Choosing BLAST options for better detection of orthologs as reciprocal best hits.** *Bioinformatics* 2008, **24**(3):319–324.
45. Shaye DS, Greenwald I: **OrthoList: A Compendium of *C. elegans* Genes with Human Orthologs.** *PLoS ONE* 2011, **6**(5):e20085.
46. Tautz D, Domazet-Lošo T: **The evolutionary origin of orphan genes.** *Nature Reviews Genetics* 2011, **12**(10):692–702.
47. Chen F, Mackey AJ, Vermunt JK, Roos DS: **Assessing Performance of Orthology Detection Strategies Applied to Eukaryotic Genomes.** *PLoS ONE* 2007, **2**(4):e383.
48. Keeling PJ, Corradi N, Morrison HG, Haag KL, Ebert D, Weiss LM, Akiyoshi DE, Tzipori S: **The Reduced Genome of the Parasitic Microsporidian *Enterocytozoon bieneusi* Lacks Genes for Core Carbon Metabolism.** *Genome biology and evolution* 2010, **2**(0):304–309.
49. Levin M, Hashimshony T, Wagner F, Yanai I: **Developmental Milestones Punctuate Gene Expression in the *Caenorhabditis* Embryo.** *Developmental Cell* 2012, **22**(5):1101–1108.
50. Phelan P: **Innexins: members of an evolutionarily conserved family of gap-junction proteins.** *Biochimica et Biophysica Acta (BBA) - Biomembranes* 2005, **1711**(2):225–245.
51. Jones AK, Sattelle DB: **Functional genomics of the nicotinic acetylcholine receptor gene family of the nematode, *Caenorhabditis elegans*.** *BioEssays* 2003, **26**:39–49.
52. Antebi A: **Nuclear hormone receptors in *C. elegans*.** In *Wormbook*, Edited by The *C. elegans* Research Community, WormBook 2006 .
53. Altun ZF, Chen B, Wang ZW, Hall DH: **High resolution map of *Caenorhabditis elegans* gap junction proteins.** *Developmental Dynamics* 2009, **238**(8):1936–1950.
54. Bowerman B: **Embryonic polarity: Protein stability in asymmetric cell division.** *Current Biology* 2000, **10**(17):R637–R641.
55. Severson AF, Bowerman B: **Myosin and the PAR proteins polarize microfilament-dependent forces that shape and position mitotic spindles in *Caenorhabditis elegans*.** *The Journal of Cell Biology* 2003, **161**:21–26.
56. Gönczy P, Rose LS: **Asymmetric cell division and axis formation in the embryo.** In *WormBook*, Edited by The *C. elegans* Research Community, WormBook 2005 .
57. Goldstein B, Macara IG: **The PAR proteins: fundamental players in animal cell polarization.** *Developmental Cell* 2007, **13**(5):609–622.

58. Cheng NN, Kirby CM, Kemphues KJ: **Control of cleavage spindle orientation in *Caenorhabditis elegans*: the role of the genes *par-2* and *par-3*.** *Genetics* 1995, **139**(2):549–559.
59. Wu JC, Rose LS: **PAR-3 and PAR-1 inhibit LET-99 localization to generate a cortical band important for spindle positioning in *Caenorhabditis elegans* embryos.** *Molecular biology of the cell* 2007, **18**(11):4470–4482.
60. Doerflinger H, Vogt N, Torres IL, Mirouse V, Koch I, Nusslein-Volhard C, St Johnston D: **Bazooka is required for polarisation of the *Drosophila* anterior-posterior axis.** *Development* 2010, **137**(10):1765–1773.
61. Goldstein B, Frisse L, Thomas W: **Embryonic axis specification in nematodes: evolution of the first step in development.** *Current Biology* 1998, **8**(3):157–160.
62. Draper BW, Mello CC, Bowerman B, Hardin J, Priess JR: **MEX-3 is a KH domain protein that regulates blastomere identity in early *C. elegans* embryos.** *Cell* 1996, **87**(2):205–216.
63. Huang N, Mootz D, Walhout A, Vidal M, Hunter CP: **MEX-3 interacting proteins link cell polarity to asymmetric gene expression in *Caenorhabditis elegans*.** *Development* 2002, **129**(3):747–759.
64. Evans TC, Hunter CP: **Translational control of maternal RNAs.** In *WormBook*, Edited by The *C. elegans* Research Community, WormBook 2005 .
65. Gomes JE, Encalada SE, Swan KA, Shelton CA, Carter JC, Bowerman B: **The maternal gene *spn-4* encodes a predicted RRM protein required for mitotic spindle orientation and cell fate patterning in early *C. elegans* embryos.** *Development* 2001, **128**(21):4301–4314.
66. Labbé JC, Goldstein B: **Embryonic development: A New SPN on cell fate specification.** *Current Biology* 2002, **12**(11):R396–398.
67. Simske JS, Hardin J: **Getting into shape: epidermal morphogenesis in *Caenorhabditis elegans* embryos.** *BioEssays* 2001, **23**:12–23.
68. Gilleard JS, McGhee JD: **Activation of Hypodermal Differentiation in the *Caenorhabditis elegans* Embryo by GATA Transcription Factors *ELT-1* and *ELT-3*.** *Molecular and Cellular Biology* 2001, **21**(7):2533–2544.
69. Sommer R: **As good as they get: cells in nematode vulva development and evolution.** *Current Opinion in Cell Biology* 2001, **13**(6):715–720.
70. Kiontke K, Barriere A, Kolotuev I, Podbilewicz B, Sommer R, Fitch DHA, Félix MA: **Trends, stasis, and drift in the evolution of nematode vulva development.** *Current biology : CB* 2007, **17**(22):1925–1937.
71. Sternberg PW: **Vulval development.** In *WormBook*, Edited by The *C. elegans* Research Community, WormBook 2005 .
72. Salser SJ, Loer CM, Kenyon C: **Multiple *HOM-C* gene interactions specify cell fates in the nematode central nervous system.** *Genes & Development* 1993, **7**(9):1714–1724.
73. Eisenmann DM, Kim SK: **Protruding vulva mutants identify novel loci and Wnt signaling factors that function during *Caenorhabditis elegans* vulva development.** *Genetics* 2000, **156**(3):1097–1116.
74. Shemer G, Podbilewicz B: **LIN-39/Hox triggers cell division and represses *EFF-1*/fusogen-dependent vulval cell fusion.** *Genes & Development* 2002, **16**(24):3136–3141.
75. Tian H, Schlager B, Xiao H, Sommer RJ: **Wnt signaling induces vulva development in the nematode *Pristionchus pacificus*.** *Current Biology* 2008, **18**(2):142–146.
76. Streit A, Kohler R, Marty T, Belfiore M, Takacs-Vellai K, Vigano MA, Schnabel R, Affolter M, Müller F: **Conserved Regulation of the *Caenorhabditis elegans* *labial/Hox1* Gene *ceh-13*.** *Developmental Biology* 2002, **242**(2):96–108.
77. Aboobaker AA, Blaxter ML: **Hox Gene Loss during Dynamic Evolution of the Nematode Cluster.** *Current Biology* 2003, **13**:37–40.
78. Aboobaker A, Blaxter M: **Hox gene evolution in nematodes: novelty conserved.** *Current Opinion in Genetics & Development* 2003, **13**:593–598.
79. Lemons D, McGinnis W: **Genomic Evolution of Hox Gene Clusters.** *Science* 2006, **313**(5795):1918–1922.
80. Bird A: **DNA methylation patterns and epigenetic memory.** *Genes & Development* 2002, **16**:6–21.

81. Gao F, Liu X, Wu XP, Wang XL, Gong D, Lu H, Xia Y, Song Y, Wang J, Du J, Liu S, Han X, Tang Y, Yang H, Jin Q, Zhang X, Liu M: **Differential DNA methylation in discrete developmental stages of the parasitic nematode *Trichinella spiralis*.** *Genome Biology* 2012, **13**(10):R100.
82. Tran RK, Zilberman D, de Bustos C, Ditt RF, Henikoff JG, Lindroth AM, Delrow J, Boyle T, Kwong S, Bryson TD, Jacobsen SE, Henikoff S: **Chromatin and siRNA pathways cooperate to maintain DNA methylation of small transposable elements in *Arabidopsis*.** *Genome Biology* 2005, **6**(11):R90.
83. Martienssen RA, Colot V: **DNA methylation and epigenetic inheritance in plants and filamentous fungi.** *Science* 2001, **293**(5532):1070–1074.
84. Chamberlin HM, Thomas JH: **The bromodomain protein LIN-49 and trithorax-related protein LIN-59 affect development and gene expression in *Caenorhabditis elegans*.** *Development* 2000, **127**(4):713–723.
85. Tingley GA, Anderson RM: **Environmental sex determination and density-dependent population regulation in the entomogenous nematode *Romanomermis culicivorax*.** *Parasitology* 1986, **92**:431–449.
86. Haag E: **The evolution of nematode sex determination: *C. elegans* as a reference point for comparative biology.** In *WormBook*, Edited by The C. elegans Research Community, WormBook 2005 .
87. Powell JR, Jow MM, Meyer BJ: **The T-box transcription factor SEA-1 is an autosomal element of the X:A signal that determines *C. elegans* sex.** *Developmental Cell* 2005, **9**(3):339–349.
88. Chu DS, Dawes HE, Lieb JD, Chan RC, Kuo AF, Meyer BJ: **A molecular link between gene-specific and chromosome-wide transcriptional repression.** *Genes & Development* 2002, **16**(7):796–805.
89. Meyer B: **X-Chromosome dosage compensation.** In *WormBook*, Edited by The C. elegans Research Community, WormBook 2005 .
90. Zarkower D: **Somatic sex determination.** In *WormBook*, Edited by The C. elegans Research Community, WormBook 2006 .
91. Kuwabara PE, Okkema PG, Kimble J: **tra-2 encodes a membrane protein and may mediate cell communication in the *Caenorhabditis elegans* sex determination pathway.** *Molecular biology of the cell* 1992, **3**(4):461–473.
92. Goodwin EB, Ellis RE: **Turning clustering loops: sex determination in *Caenorhabditis elegans*.** *Current Biology* 2002, **12**(3):R111–20.
93. Baldi C, Cho S, Ellis RE: **Mutations in two independent pathways are sufficient to create hermaphroditic nematodes.** *Science* 2009, **326**(5955):1002–1005.
94. Mohamed AK, Burr C, Burr AHJ: **Unique Two-Photoreceptor Scanning Eye of the Nematode *Mermis nigrescens*.** *The Biological Bulletin* 2007, **212**(3):206–221.
95. Robinson AF, Baker GL, Heald CM: **Transverse Phototaxis by Juveniles of *Agamermis* sp. and *Hexamermis* sp.** *The Journal of Parasitology* 1990, **76**(2):147–152.
96. Shamseldeen MM, Platzer EG, Gaugler R: **Role of the surface coat of *Romanomermis culicivorax* in immune evasion.** *Nematology* 2007, **9**:17–24.
97. Tang K, Xie X, Park JI, Jamrich M, Tsai S, Tsai MJ: **COUP-TFs regulate eye development by controlling factors essential for optic vesicle morphogenesis.** *Development* 2010, **137**(5):725–734.
98. Burr AHJ, Wagar D, Sidhu P: **Ocellar pigmentation and phototaxis in the nematode *Mermis nigrescens*: changes during development.** *The Journal of Experimental Biology* 2000, **203**:1341–1350.
99. Burr AH, Hunt P, Wagar DR, Dewilde S, Blaxter ML, Vanfleteren JR, Moens L: **A hemoglobin with an optical function.** *The Journal of biological chemistry* 2000, **275**(7):4810–4815.
100. Hunt P, McNally J, Barris W: **Duplication and divergence: the evolution of nematode globins.** *Journal Of Nematology* 2009, **41**:35–51.
101. Fairbank PD, Lee C, Ellis A, Hildebrand JD, Gross JM, Wallingford JB: **Shroom2 (APXL) regulates melanosome biogenesis and localization in the retinal pigment epithelium.** *Development* 2006, **133**(20):4109–4118.
102. De Ley P: **A quick tour of nematode diversity and the backbone of nematode phylogeny.** In *Worm-Book*, Edited by The C. elegans Research Community, WormBook 2006 .

103. Martin M: **Cutadapt** removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 2011, **17**.
104. CLCbio: **White paper on de novo assembly in CLC Assembly Cell**. Tech. rep.
105. Kumar S, Blaxter ML: **Simultaneous genome sequencing of symbionts and their hosts**. *Symbiosis* 2012, **55**(3):119–126.
106. Parra G, Bradnam K, Korf I: **CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes**. *Bioinformatics* 2007, **23**(9):1061–1067.
107. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W: **Scaffolding pre-assembled contigs using SS-PACE**. *Bioinformatics* 2011, **27**(4):578–579.
108. Li W, Godzik A: **Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences**. *Bioinformatics* 2006, **22**(13):1658–1659.
109. Smit AFA, Hubley R, Green P: **RepeatMasker Open-3.0. 1996-2010**. <http://www.repeatmasker.org>.
110. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J: **Repbase Update, a database of eukaryotic repetitive elements**. *Cytogenetic and genome research* 2005, **110**(1-4):462–467.
111. Volfovsky N, Haas BJ, Salzberg SL: **A clustering method for repeat analysis in DNA sequences**. *Genome Biology* 2001, **2**(8):research0027.1—0027.11.
112. Bao Z, Eddy SR: **Automated De Novo Identification of Repeat Sequence Families in Sequenced Genomes**. *Genome Research* 2002, **12**(8):1269–1276.
113. Price AL, Jones NC, Pevzner PA: **De novo identification of repeat families in large genomes**. *Bioinformatics* 2005, **21**(Suppl 1):i351–i358.
114. Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, Holt C, Sánchez Alvarado A, Yandell M: **MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes**. *Genome Research* 2008, **18**:188–196.
115. Parkinson J, Mitreva M, Whitton C, Thomson M: **A transcriptomic analysis of the phylum Nematoda**. *Nature Genetics* 2004, **36**:1259–1267.
116. Stanke M, Waack S: **Gene prediction with a hidden Markov model and a new intron submodel**. *Bioinformatics* 2003, **19**(2):ii215–ii225.
117. Gremme G, Brendel V, Sparks ME, Kurtz S: **Engineering a software tool for gene structure prediction in higher organisms**. *Information and Software Technology* 2005, **47**(15):965–978.
118. Supek F, Vlahovicek K: **INCA: synonymous codon usage analysis and clustering by means of self-organizing map**. *Bioinformatics* 2004, **20**(14):2329–2330.
119. Cutter AD, Wasmuth JD, Blaxter ML: **The evolution of biased codon and amino acid usage in nematode genomes**. *Molecular Biology And Evolution* 2006, **23**(12):2303–2315.
120. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M: **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research**. *Bioinformatics* 2005, **21**(18):3674–3676.
121. The Gene Ontology Consortium: **Gene Ontology: tool for the unification of biology**. *Nature Genetics* 2000, **25**:25–29.
122. Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R: **InterProScan: protein domains identifier**. *Nucleic Acids Research* 2005, **33**(Web Server):W116–W120.
123. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy S, Bateman A: **Rfam: annotating non-coding RNAs in complete genomes**. *Nucleic Acids Research* 2004, **33**(Database issue):D121–D124.
124. Li L, Stoeckert CJ, Roos DS: **OrthoMCL: identification of ortholog groups for eukaryotic genomes**. *Genome Research* 2003, **13**(9):2178–2189.
125. van Dongen S: **A Cluster algorithm for graphs**. *Report - Information systems* 2000, (10):1–40.
126. van Dongen S: **Graph Clustering by Flow Simulation**. *PhD thesis*, University of Utrecht 2000.
127. Altschul SF, Gish W, Miller W, Myers EW: **Basic local alignment search tool**. *Journal of Molecular Biology* 1990, **215**:403–410.

128. Seydoux G, Fire A: **Whole-mount in situ hybridization for the detection of RNA in *Caenorhabditis elegans* embryos.** *Methods in cell biology* 1995, **48**:323–337.
129. Wang J, Mitreva M, Berriman M, Thorne A, Magrini V, Koutsovoulos G, Kumar S, Blaxter ML, Davis RE: **Silencing of Germline-Expressed Genes by DNA Elimination in Somatic Cells.** *Developmental Cell* 2012, **23**(5):1072–1080.

PREPRINT ONLY

Figure 1

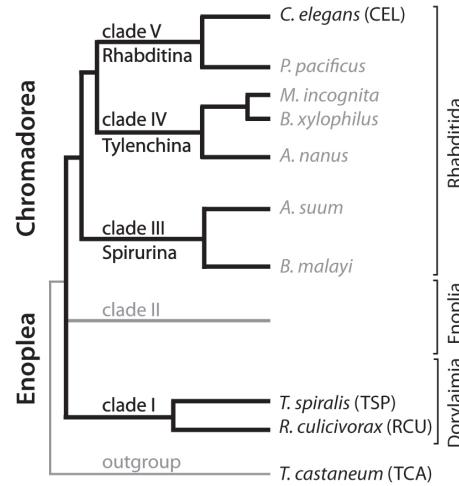


Figure 1: **A simplified phylogenetic tree of the phylum Nematoda.** The phylogeny, simplified from [1, 12], emphasises the position of the main study species *R. culicivora*, *T. spiralis* and *C. elegans*. The phylogenetic placements of species from Table 2 are given in grey. Currently no genomic data are available for Enoplea (Clade II). The order of branching of the basal nodes of Nematoda is unresolved.

Figure 2

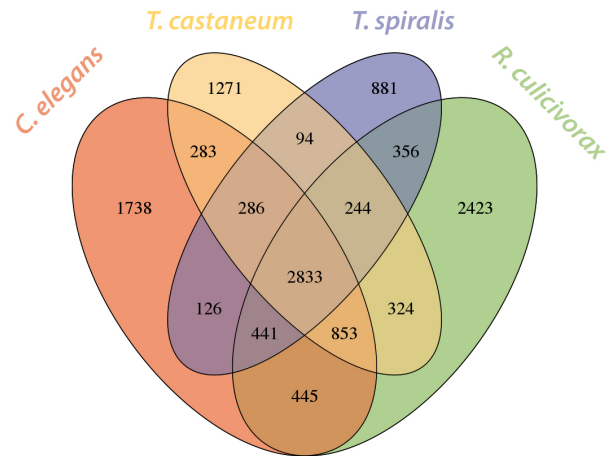


Figure 2: **Clusters of homologous proteins.** Shared and species-unique clusters of homologous proteins from a comparison of the proteomes of *Romanomermis culicivorax*, *Trichinella spiralis*, *Caenorhabditis elegans* and *Tribolium castaneum* using OrthoMCL.

Figure 3

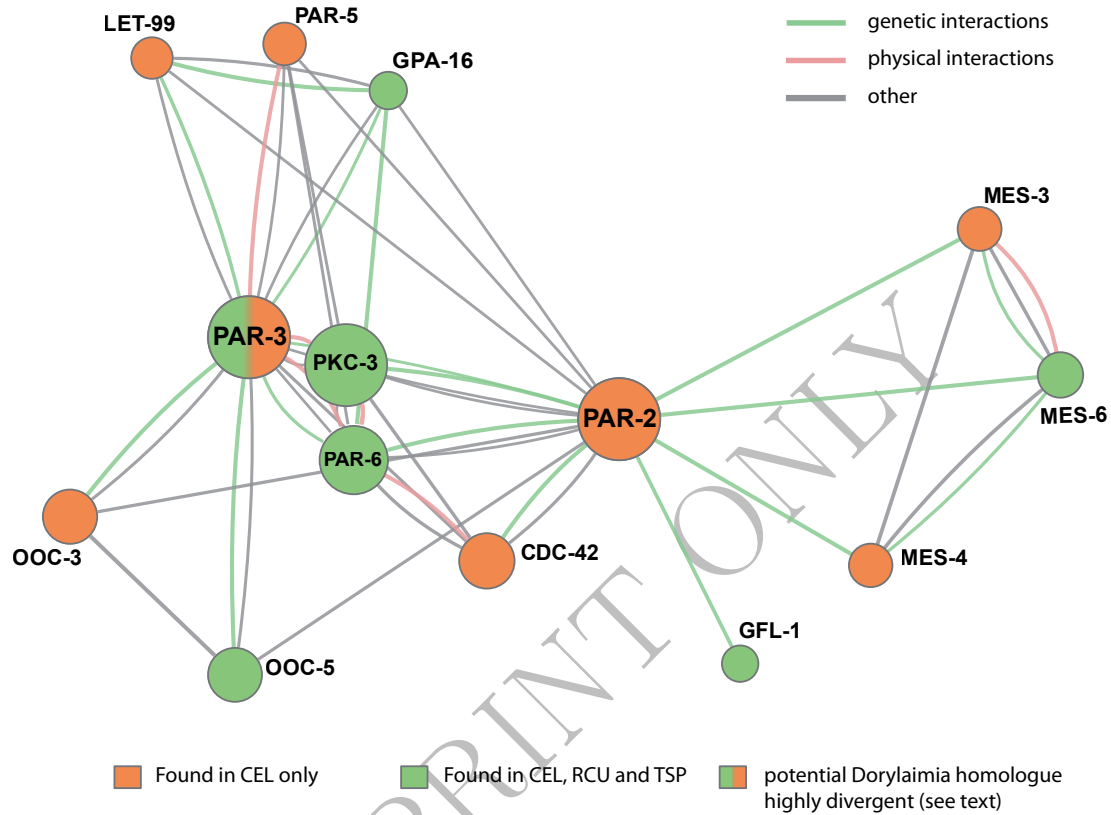


Figure 3: The network of proteins interacting with PAR-2 and PAR-3 in *Caenorhabditis elegans* and their orthologues in *Romanomermis culicivorax* and *Trichinella spiralis*. The network cartoon is based on the core polarity pathway extracted from WormBase, derived from both genetic and physical interactions. PAR-2 was missing from the dorylaim nematodes, as were the directly connected *mes-3* and *mes-4* genes. The *R. culicivorax* PAR-3-like proteins was not retrieved as an orthologue of *C. elegans* and *T. spiralis* PAR-3 proteins, but was identified employing sensitive sequence similarity search. See Table 3 for additional proteins interacting with PAR proteins and their presence-absence patterns.

Figure 4

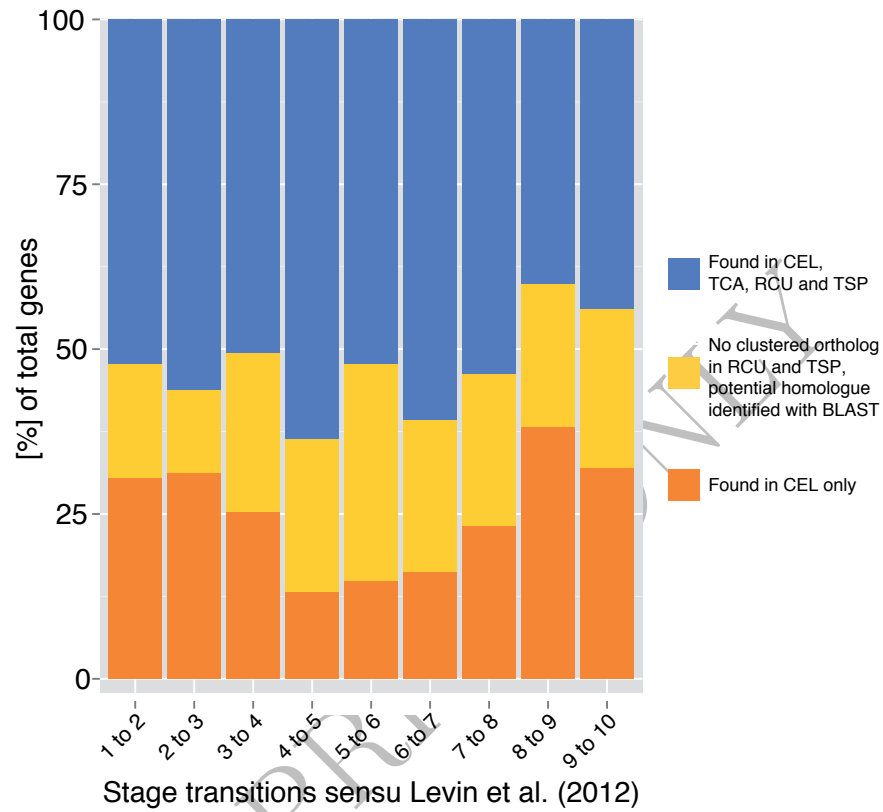


Figure 4: Many genes that are developmentally important in *Caenorhabditis elegans* were not present in *Romanomermis culicivorax* or *Trichinella spiralis*. *R. culicivorax* and *T. spiralis* orthologues of the 1,725 genes identified as important in embryogenesis in an analysis of gene expression in *Caenorhabditis* species [49] were sought. For each embryonic stage (1-10) in *C. elegans* we calculated the proportion of these genes that were apparently unique to the genus *Caenorhabditis*.

Figure 5

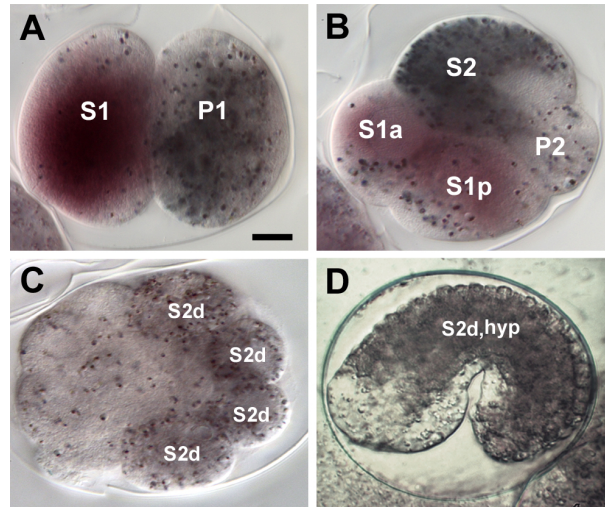


Figure 5: *In situ* hybridisation revealing the pattern of distribution of *mex-3* mRNA in early embryos of *Romanomermis culicivorax*. We used the *R. culicivorax mex-3* gene to prove application of the *in situ* technique in this species and investigate the patterns of segregation of this maternal RNA in early development. The *R. culicivorax mex-3* expression pattern is similar to that of *C. elegans* [62]. *R. culicivorax* embryos contain dark pigment granules that are asymmetrically segregated in development. (A) At the 2-cell stage, maternal *mex-3* mRNA is detected in the S1 blastomere. The cytoplasmic pigment granules are predominantly in the P1 blastomere. (B) At the 4-cell stage, *mex-3* mRNA is detected in daughters of anterior S1 cell. Cytoplasmic pigment granules are predominantly in the S2 blastomere. (C) At a later stage (>20 cells), *mex-3* mRNA is absent. The pigment granules are found in descendants of S2 (S2d). (D) During early morphogenesis, the pigment granules are found in S2 descendants forming hypodermis, (S2d, hyp). (A-C) fixed embryos; (D) live embryo. Bar 10 μ m. Orientation: anterior left.

Tables

Table 1- Assembly and annotation statistics

Metric	Result
Contigs >100bp span	267,342,457bp
Scaffolds >500bp span	322,765,761bp
Num. contigs/scaffolds	62,537
N50 contigs/scaffolds >500bp	17,632 bp
N50 scaffolds >500bp	29,995bp
Max contig length	28,847bp
Max scaffold length	201,054bp
Mean transcript length	593bp
Mean protein length	190aa
MAKER Augustus predictions	12,026 proteins
MAKER SNAP predictions	36,145 proteins
Num. ESTs (isogroups)	22,418 ESTs
Mean EST length	330bp
80% BLAT sequence coverage	21,204 ESTs
CEGMA compl. completeness	75.40%
CEGMA Group 1 part. compl.	81.82%
CEGMA Group 2 part. compl.	91.07%
CEGMA Group 3 part. compl.	91.80%
CEGMA Group 4 part. compl.	95.38%

Table 2 - Genome statistics

Repeat content of different nematode genomes appears not to be directly correlated with genome size. Re-calculation in selected genomes shows little deviance from published data (in parentheses)* and thus indicates the validity of our inference for *R. culicivora*.

*For *B. xylophilus* and *M. incognita* only reference data is given as the same programs were used for initial inference (see references); *A. suum* not re-calculated.

Species	Approximate [#] genome size	Estimated Repeat content	Median [†] exon length	Median [†] intron length	GC content	Source
<i>C.elegans</i>	100Mb	17% (16.5%)	145bp	69bp	38%	[17, 18]
<i>P. pacificus</i>	165Mbp	15.3% (17%)	85bp	141bp	42%	[20, 25]
<i>A. suum</i>	334Mb	4.4%	144bp	907bp	37.9%	[21, 129]
<i>B. malayi</i>	95Mb	16.5% (15%)	140bp	219bp	30%	[22]
<i>B. xylophilus</i>	69Mb	22.5%	183bp	69bp	40%	[25]
<i>M. incognita</i>	~200Mb	36.7%	136bp	82bp	31%	[24]
<i>T. spiralis</i>	63Mb	19.8% (18%)	128bp	283bp	34%	[26]
<i>R. culicivora</i>	>270Mb	48.2%	161bp	405bp	36%	this work

[#] *M. incognita* genome size given as 86Mbp in [24] has been re-estimated to about 200Mbp (E. Danchin pers. comm.).

[†]Median lengths for *A. suum* and *T. spiralis* were calculated in this work as these data are not given in the cited publications.

Table 3 Presence and absence of selected *C. elegans* proteins in *Dorylaimia*

Protein	<i>T. spiralis</i>	<i>R. culicivora</i>
Early asymmetry		
CDC-42	+	+
PKC-3	+	+
GPR-1	+	+
GPR-2	+	+
PAR-6	+	+
MES-6	+	+
MES-3	-	-
MES-4	-	-
GFL-1	+	+
LET-70	+	+
Axis formation		
NUM-1	+	+
ZIM-1	-	-
MES-2	-	-
POS-1	-	-
SMA-6	+	+
SET-2	-	-
UBC-18	+	+
LET-99	-	-
OOC-3	-	-
OOC-5	+	+
GPA-16	+	+
PAR-5	-	-
ATX-2	-	-
MEX-5	-	-
MEX-6	-	-
UNC-120	-	-
NOS-2	-	-
OMA-1	-	-
RME-2	+	+
SPN-4	-	-
Sex determination		
XOL-1	-	-
HER-1	-	-
SEX-1	+	+
FOX-1	+	+
SDC-1	-	-
SDC-2	-	-
SDC-3	-	-
TRA-2	-	-
FEM-1	+	+
FEM-2	+	+
Hypodermis and vulva formation		
AFF-1	-	-
BAR-1	-	-
CEH-2	-	-
CEH-27	-	-
GRL-15	-	-
INX-5	-	-
LIN-1	-	-
PEB-1	-	-
ELT-3	-	-
ELT-1	+	+
SMA-3	-	-
SMA-5	-	-

Supplementary Files

These will be available through the main author upon personal request in the preprint phase.

Supplementary file 1 — Supplementary Figures and Tables

Supplementary file 2 — Fisher's exact test data

GO terms enriched in a set of protein clusters shared between *Dorylaimia* in comparison to (i) *C. elegans* and (ii) *T. castaneum* proteomes.

Supplementary file 3 — Codon usage in *R. culicivox*

Codon usage data.

Supplementary file 4 — Levin data

Genes identified as being differentially expressed in *Caenorhabditis* development by Levin et al. [49].

Supplementary file 5 — Analysis of orthoMCL output by BLAST+

BLAST+ results for specific *C. elegans* proteins not found in a cluster with *Dorylaimia* proteins.

PREPRINT ONLY